論文97-34C-5-8

결합정보를 이용한 명사 및 접사 추출

(Noun and Affix Extraction using Conjunctive Information)

徐 昌 德 *, 林 寅 七 *

(Changduck Suh and In-Chil Lim)

요 약

본 논문에서는 형태소분석 및 구문분석의 결과를 이용하여 색인어를 추출하는 시스템을 구축하기 위해 결합정보를 이용한 명사 및 접사 추출 방법을 제안한다. 다른 언어와 달리 독특한 띄어쓰기 규칙을 가지고 있는 한글 입력문장으로부터 도출한 결합정보를 이용하면 최소한의 비용으로 다품사의 품사후보 수를 줄일 수 있다. 또한 하나의 어절이 줄바꿈문자에 의해 강제 분리된 경우 이를 결합정보의 관점에서 해결한다. 제안한 알고리듬은 형태소분석하는 과정을 통해 그 효율성을 보인다.

Abstract

This paper proposes noun and affix extraction methods using conjunctive information for making an automatic indexing system through morphological analysis and syntactic analysis. The Korean language has a peculiar spacing words rule, which is different from other languages, and the conjunctive information, which is extracted from the rule, can reduce the number of multiple parts of speech at a minimum cost. The proposed algorithms also solve the problem that one word is seperated by newline character. We show efficiency of the proposed algorithms through the process of morphological analyzing.

I. 서 론

많은 자료를 DB화하여 원하는 문서를 검색하고자 할 때는 문서별 키워드 색인작업이 이루어져야 한다. 과거 이러한 작업은 사람에 의해서 이루어져 왔으나 많은 노동력을 필요로 하므로 문헌을 전자문서화 하여 키워드 검색이 가능하도록 하기 위해서는 컴퓨터에 의한 자동색인(automatic indexing)이 일반적이다. 해당 문헌에 대해 자동색인한 결과가 전문가의 수작업에 의한 색인 결과와 비교해 얼마나 근접할 수 있느냐가 관건인데 이를 위한 척도로 재현율(recall rate)과 정확도(precision rate)를 가지고 평가한다[1].

* 正會員, 漢陽大學校 電子工學科

(Dept. of Electronics Eng., Hanyang University)

接受日字:1996年11月26日, 수정완료일:1997年5月2日

원하는 문서를 많이 검색해주는 비율인 재현율보다는 검색된 문서가 실제 얼마나 맞는가를 나타내는 정도인 정확도에 더 비중을 두는 쪽으로 연구가 이루어지고 있으며 문서로부터 색인어를 자동 추출하는 방법 중에는 어휘사전을 이용하지 않고 조사, 어미 등만 수록한 기능어 사전을 이용해 명사를 추출하는 방법이 있다. 이러한 방법은 어휘사전이 필요 없다는 장점이 있으나 명사추출이 불완전해 오분석을 유발하고 복합명사 처리에 어려움이 있다.[3]

이에 반해 어휘사전을 이용하면 형태소분석이 가능해 언어규칙과 의미망(semantic network)을 이용하여 구문분석, 의미분석 등 언어학적 방법으로 접근할 수 있다. 형태소분석이나 구문분석까지 거친 후 추출된 색인어들은 일반적으로 출현 빈도수를 중요도의 척도로 삼으며 색인어가 문서 내에 위치한 위치정보, 불용어

사전에 의한 불용어 제거, 시소러스를 이용한 어의확장 등 자동색인에 관련된 많은 기법들을 적용하여 한 문서에 대한 색인어들간의 가중치를 결정한다. 이렇게 하여 최종적으로 선택된 색인어들은 질의시 문서간 순위(ranking)를 결정하는데 이용된다.[1,2] 따라서 문서로부터 명사를 추출하는 방법 또한 단순한 규칙이나 통계적인 방법보다는 언어학적 분석으로 명사를 추출해 내어 색인어 선정에 이용하는 방법이 많이 연구되고 있는 실정이다[12].

구문분석, 의미분석 쪽으로 갈수록 보다 완벽한 분석이 가능해 정확한 색인어 추출을 기대할 수 있지만 의미분석에 대한 연구는 아직도 미진한 상태이며 완벽한 구문분석 또한 기대하기 힘든 실정이다. 그리고 구문분석, 의미분석에 기울인 노력, 시스템의 복잡성, 처리시간 증가에 대해 색인어 추출결과의 정확성이 생각만큼 향상되지 않기 때문에 일반적으로 형태소분석이나 대략적인 구문분석 단계까지만 수행하는 경우가 많다[2].

형태소분석시 발생하는 모호성 중에는 하나의 형태소가 여러 품사를 가지는 형태소 자체의 품사 모호성과 하나의 어절이 여러 형태소로 분해되면서 발생하는 구조적 모호성 등이 있다. 전자에 관한 연구로는 구문분석 단계에서의 과다한 부담을 줄이기 위해 구문분석의 전처리 과정으로 사용되는 품사 태깅이 있으며[8] 후자에 관해서는 특히 색인어 추출과 관련해 복합명사를 분해하기 위한 연구가 있다.

색인어 추출 대상은 주로 명사로, 복합명사의 분해 및 명사구로부터 명사합성에 관한 연구[3-7]가 이루어져 왔으며 최근 미등록어, 고유명사, 인명, 지명, 한자 색인어 추출 등에 관한 연구도 함께 이루어지고 있다. 그러나 미등록어를 처리하기에 곤란한 품사태거는 시스템마다 설정된 수십 가지의 품사분류가 서로 다르고 세부 분류기준이 마련되어 있지 않으며 학습시키기 위한 태깅된 코퍼스(corpus)나 균일한 원시 코퍼스를 구축하기에 많은 비용과 처리시간을 필요로 한다[8]. 품사만으로 구성된 접속정보 또한 중의성 해결을 위해 더 세부적인 품사를 사용하거나 품사와 함께 해당 형태소까지 규칙에 포함시키고 있다[9]. 이처럼 품사설정의 정확도를 어느 정도 높이기 위해서는 많은 규칙 또는 대량의 코퍼스 구축이 필수적이며 많은 비용과 처리시간을 필요로 하므로 명사 추출이 주목적인 자동색인 시스템에서는 최장일치법을 이용한 형태소분석 방법이 많이 사용된다.

본 논문은 명사 및 접사 추출을 대상으로 하는 색인 시스템을 위한 형태소분석기를 구현한다. 기계번역이나 자연언어처리 자체를 위한 형태소분석기나 품사태거가 아니라 비정형화된 전문(full text)으로부터 색인어를 추출하기 위한 형태소분석기를 구현하는데 있어, 최장일치법 적용시 발생하는 많은 품사들을 결합정보를 이용한 최소비용으로 후보 수를 줄이며 결합정보의 관점에서 접사와 줄바꿈문자를 처리하는 방법에 대해 기술한다.

제안하는 방법은 복잡한 규칙이나 품사의 세부분류, 방대한 코퍼스, 의미사전 구축 등에 따르는 문제점을 피할 수 있으며 결합정보는 형태소분석 단계에서 형태소의 다품사를 여과(filtering)하기 위한 매우 간단하고도 효율적인 방법으로 빠른 색인 시스템 구축시 또는 부담이 큰 구문분석에 앞서 사용될 수 있다. 또한 색인어와 질의문의 접사처리 여부는 질의시 재현율과 정확도 모두에 영향을 미치지만 최근 접미사의 사전정보 구축을 위한 작업으로 언어학적 관점에서 분석하고 있는 남윤진의 논문[10]을 제외하고는 미진한 실정이다. 본 논문에서는 명사는 물론 결합정보의 관점에서 접사를 처리하며, OCR에 의해 전자문서화된 문서의 매 줄 끝마다 포함된 줄바꿈문자로 인해 분석실패나 오분석이 유발되는 문제도 처리하도록 한다.

Ⅱ. 결합정보를 이용한 형태소분석기

형태소분석시 분석 성공률을 높여 여러 개의 분석후보를 내게 됨으로써 구문분석에 많은 부담을 주는 방법보다는 분석결과가 얼마나 정확한가 하는 것이 주된 관심사로 대두되고 있다[10]. 본 논문에서는 색인어 추출을 위주로 한 형태소분석기를 구현하는데 있어 가급적 적은 분석후보를 발생하되 정확한 분석이 되도록 결합정보를 이용한 방법을 제안한다. 색인어 추출을 위해 명사는 물론 접사 추출에 중점을 두며 어절분리, 줄바꿈문자 처리도 결합정보를 이용한다.

1. 결합정보

한국어 문장의 띄어쓰기는 매우 독특한 규칙을 갖는다. 영어는 단어별로 모두 띄어쓰며 일어는 모두 붙여쓰지만 한글은 각 형태소들이 일정한 규칙에 의해 붙여쓰거나 띄어 써야만 하는 경우와 양쪽 모두 가능한 경우가 있다. 결합정보를 이용한 방법은 한국어 문장으로부터 이러한 띄어쓰기 정보를 자동 획득하여 중의성

을 해결하고자 하는 시도로서, 형태소가 결합하는데 따르는 복잡한 통사적, 의미적 제약이 아닌 단순히 하나의 형태소가 앞(좌측) 형태소와 결합시 공백 없이 인접하여 와야만 하는지('+') 아닌지('-') 혹은 어느 쪽도 상관없이 사용되는지('*')를 파악하는 것이다[11].

결합정보는 임의의 형태소 m_i가 앞(좌측)형태소 m_{i-1}과 어떻게 결합하는지를 나타내는 좌결합정보와 뒤(우측)형태소 m_{i+1}과의 우결합정보로 나눌 수 있지만 본 논문에서는 특별한 언급이 없는 한 결합정보는 좌결합정보를 뜻한다.

일반적으로 조사, 어미와 같은 기능어는 다른 품사의 형태소와 동형이의어 관계에 있는 경우가 많으므로 좌결합정보는 이들을 판별하는데 도움이 된다. 즉, 기능어가 주로 앞 형태소와 붙여쓰므로 띄어쓰는 형태소와 동형이의어 관계에 있을 때 문장으로부터 획득한 결합정보로 어느 쪽인지를 쉽게 판별할 수 있다.

이에 반해 뒤 형태소와의 결합형태는 현 형태소의 우결합정보에 의한다기보다는 뒤 형태소에 따라 좌우되는 경우가 대부분이며 부사나 관형사 등은 앞뒤(좌우) 모두 띄어 쓰므로 우결합정보도 '-'라고 할 수 있지만 부사나 관형사 중 다품사인 경우는 많지 않으며 좌결합정보를 거쳐 우결합정보까지 사용해 최종적으로 하나의 후보가 선택될 수 있는 경우는 더욱 적기 때문에 우결합정보는 좌결합정보에 비해 유용성이 떨어진다.

다만 접두사의 경우는 색인어 추출과 직접적인 관련이 있고 뒤 형태소와 반드시 붙여쓰므로 우결합정보가 필요하다. 만약 형태소 X가 접두사를 포함하는 다품사이고 Y는 명사라고 했을 때 'ъXъY'(ъ는 공백)형태로 결합되었다면 X는 X의 좌결합정보만 가지고는 접두사를 걸러내지 못한다. 따라서 좌결합정보를 이용한 형태소분석이 끝난 후 남은 분석후보를 다시 한번 걸러내기 위한 여과과정에서 접두사, 부사, 관형사 등에 한해 우결합정보와 간단한 품사결합정보만을 사용한다. 일반적으로 결합정보는 품사에 의해 좌우되지만 띄어쓰기와 붙여쓰기 모두 허용되는 경우도 있고 실제 규칙과는 다르게 사용되는 경우가 많은 형태소가 있다. 특히, 1음절로 된 기능어는 다른 글자와 붙여쓰는 경향이 강한데 예로 '등(불완전명사), 및(부사), 전(관형사)' 등이 있다. 띄어쓰기 오류를 허용한 문서의 경우 같은 품사라 할지라도 특정 형태소의 경우 문서, 위치, 저자에 따라 다르게 사용될 수 있으므로 본 시스템에서는 좌결합정보를 품사단위가 아닌 형태소별로 어휘사전에 모두 기호(+,-,*)로 기록한다. 그러나 우결합정보는 앞서 언급한대로 적용대상이 되는 품사가 제한적이고 결합정보가 고정적이므로 어휘사전에 따로 기록하지 않고 품사정보로부터 유추하도록 한다.

결합정보를 어휘사전에 기록하는데는 특별한 노력이 들지 않는다. 먼저 각 품사별로 결합정보를 일률적으로 부여하는데 이 작업은 사전 관리 시스템에 의해 간단히 이루어지며 띄어쓰기 규칙 오류를 허용하는 색인기를 만들고자 한다면 틀리기 쉬운 형태소에 한해 개별적으로 '*'을 부여하면 된다.

다음 표 1은 본 논문에서 사용되는 결합정보를 나타내는 결합기호이다.

표 1. 결합기호
Table 1. Conjunctive symbols.

기호	의 미
+	공백 없이 형태소간 결합
-	공백다음에 결합
*	공백이 올 수도 안 올 수도 있다.
~	결합하여 하나의 어절로 간주

어휘사전에 기록된 결합기호로는 '+, -, *' 3가지만 있으며 입력문장으로부터 도출될 수 있는 기호는 '~'가 추가된 4가지이다. '~'는 입력문장에서 줄바꿈문자에 의해 '*'가 1차적으로 부여되었을 때 앞 어절과 반드시 결합해야한다고 판단되는 경우 '*'가 '~'로 바뀐다.

결합기호는 인덱스내 참조횟수를 기록하는 필드의 2bit만을 할당해 쓰면 되므로 주메모리에 상주시키기 위한 부담이 없으며 어휘사전에서 읽어오기전 입력문장으로부터 획득한 결합정보와 인덱스의 결합정보를 비교해 걸러져야할 동형이의어를 알 수 있으므로 불필요한 형태소를 읽어오지 않아 디스크 참조 횟수를 줄일 수 있다.

그림 1은 어휘사전에 대한 인덱스 테이블 구조를 나타낸 것으로 addr은 기능어와 기타 자주 사용되는 형태소를 따로 모아 놓은 메모리 상주 캐시테이블(cache table)의 인덱스 번호를 뜻한다. 단 0은 캐시테이블에 없고 하드디스크에 있다는 뜻이며 255는 불용어(stop-word)임을 뜻한다.

결합정보를 이용하여 형태소분석시 다품사의 후보수를 줄일 수 있는 정도는 형태소 k개가 임의의 문서에서 걸러질 평균확률로 식(1)과 같다.

```
typedef struct indexrec {
    unsigned char *kor;  //한국어 형태소
    unsigned char ci_m;  //결합정보(2bit) + 참조횟수(6bit)
    unsigned char addr;  //0:hard, 1-254:RAM, 255(-1):불용어
} IndexRec;
```

그림 1. 인덱스 테이블 구조
Fig. 1. The structure of index table.

$$\frac{\sum_{i=1}^{k}(P^{+}\cdot r_{i+} + P^{-}\cdot r_{i-})P_i}{k} \quad (1)$$

다음은 식(1)에 관련된 기호정의이다.

k : 유일한 형태소 수

P_i : 형태소 i가 임의의 문서에서 사용될 확률

P^c : 문서에서 사용된 형태소 i가 결합기호 c의 형태로 사용될 확률

($c \in \{+,-,*\}$, $0 \le P^c \le 1$, $P^{+}+P^{-}+P^{*}=1$)

r_{ic} : 형태소 i가 결합기호 c에 의해 걸러지는 비율($r_{ic} = n_{xc}/n_i$, $0 \le r_{ic} < 1$)

(c='*'인 경우 결합정보를 적용하지 않은 경우와 같으므로 $r_{i*} = 0$)

$P^c \cdot r_{ic}$: 문서에서 사용된 형태소 i가 결합기호 c의 형태로 사용되어 걸러질 확률

$= P^{+}\cdot r_{i+} + P^{-}\cdot r_{i-} + P^{*}\cdot r_{i*} = P^{+}\cdot r_{i+} + P^{-}\cdot r_{i-}$

n_i : 어휘사전에 수록된 형태소 i의 다품사(레코드) 수 ($n_i = n_{xc} + n_{rc} = n^{+} + n^{-} + n^{*}$)

n_{xc} : 형태소 i가 결합기호 c에 의해 걸러지는 품사 수($n_{xc} \in \{n^{+},n^{-},0\}$, $n_{rc} = n_i - n_{xc}$)

n^c : n_i중 결합기호 c를 갖는 레코드 수($n^c \in \{n^{+},n^{-},n^{*}\}$)

n : 형태소 k개의 총 다품사 수(= 어휘사전에 수록된 수 = $\sum_{i=1}^{k} n_i$)

표 2. 각 결합기호에 의한 n_{xc}, n_{rc}, r_{ic}
Table 2. The n_{xc}, n_{rc}, and r_{ic} by each conjunctive symbol.

	n_{xc}	n_{rc}	r_{ic}
c='+'	n^{-}	$n^{*} + n^{+}$	n^{-} / n_i
c='-'	n^{+}	$n^{*} + n^{-}$	n^{+} / n_i
c='*'	0	n_i	0

임의의 형태소 i가 특정 문서에서 '+,-,*'의 형태로 각각 사용되었다면 각 결합기호에 의해 걸러지는 수 n_{xc}, 걸러지고 남는 수 n_{rc}와 걸러지는 비율 r_{ic}는 표 2와 같다.

결합기호 '*'는 품사 후보수를 줄이는데 아무런 도움이 되지를 못하지만 띄어쓰기 문법에 모두 맞추어 쓴 문서를 분석대상으로 한다면 명사를 제외한 나머지 품사의 경우 '*'를 '+'나 '-'로 바꿀 수 있으므로 효율이 높아진다. 그러나 어느 정도 효율이 떨어지는 것을 감수하고 띄어쓰기 오류를 허용하는 형태소분석기를 만들고자 한다면 띄어쓰기 오류가 종종 발생하는 형태소에 대해 '*'를 부여하면 된다.

결합정보에 의해 걸러지는 품사에 대한 타당성 여부는 입력문장의 띄어쓰기 형태에 따라 좌우되는데 다음과 같이 5가지 경우가 있다.

① 띄어쓰기 규칙이 지켜진 경우
② 규칙을 잘 몰라 붙이거나 띄어 쓰는 경우
③ 붙여써야 하지만 줄바꿈문자에 의해 분리된 경우
④ 띄어써야 하지만 붙여쓴 경우
⑤ 붙여써야 하지만 공백문자로 분리된 경우

```
int LoadMorph(morphem, cis, i)
char *morphem          //형태소
char cis               //문장으로부터의 결합정보
unsigned long int i  //읽어올 형태소의 다품사 k개 중 맨위 레코드 위치
{
    unsigned char cid //어휘사전으로부터의 결합정보
    Morph *cm
    동적노드생성(cm)
    WHILE morphem = mindex[i].kor     //다품사 k개 내에서 검사
①     cid = mindex[i].ci_m & 0xC0  //cid 도출
②     IF cis='*' OR (cis='+' AND cid=0x40) OR (cis='-' AND cid=0x80)
        ReadRec(i,cm)    //cis와 cid를 비교하여 관련있는 레코드만 읽음
      ENDIF
      i++
    ENDWHILE
    링크설정
}
```

그림 2. 결합정보를 이용한 다품사 여과 알고리듬
Fig. 2. The algorithm of filtering multiple parts of speech using the conjunctive information.

대부분 ①이나 ②에 해당되는데 OCR에 의한 문서의 경우는 매 줄 끝마다 발생하므로 ③의 빈도수가 높게 된다. ①의 경우 걸러지는 품사에 대한 타당성은 100%

이며 ②의 경우 틀리기 쉬운 경우의 형태소에 한해 '*'을 부여하면 잘못 제거하지는 않는다. 다만 다품사 감소율이 줄어든다. ③은 줄바꿈문자 처리로 해결하면 되며 ④의 경우는 신문에서 발견되는 형태로 분석 성공하지만 결합정보만으로는 여전히 오분석할 확률이 존재한다. ⑤는 제목 등을 쓸 때 이외에는 거의 발견되지 않는 형태로 오분석 또는 분석 실패하게 된다. 본 논문에서는 ①-③의 경우를 위주로 처리한다.

그림 2는 문장으로부터 도출된 결합정보를 어휘사전의 결합정보와 비교해 관련 없는 다품사 형태소를 사전에서 읽어오지 않도록 하는 알고리듬이다.

이 중 ①②문이 결합정보를 적용하는 부분으로 ①에서 bit연산(bitwise AND, C언어의 &)과 ②에서 IF문에서의 논리연산(logical OR, AND) 및 비교에 걸리는 약간의 시간만 소비하면 다품사 수를 줄일 수 있어 어휘사전 참조횟수와 구문분석에 대한 부담 및 처리시간을 줄일 수 있다. 그림 3은 메모리 인덱스로부터의 결합정보 cid를 추출해 입력문장으로부터 획득한 결합정보 cis와 비교해 사전에서 읽어 올 것인지를 결정하기 위한 과정을 보이고 있다.

2. 어절분리

형태소분석은 일반적으로 어절단위로 이루어지며 어절은 공백문자로 구분된다. 그러나 서로 다른 언어가 붙어서 하나의 어절로 구성된 경우 그대로 분석을 시도하면 사전참조횟수가 늘어나므로 본 형태소분석기는 하나의 어절이라 할지라도 서로 다른 언어와의 경계점 역시 어절을 구분하는 단위로 본다. 즉, 문장분석대상은 한글이지만 문서에는 다양한 언어가 포함되어 있으므로 이를 한글, 한자, 영어, 숫자, 심벌(1byte, 2byte), 공백문자로 구분하여 어절분리에 이용한다. 공백문자 다음에 오는 어절은 결합기호로 '-', 즉 어절의 처음 형태소에 '-'가 부여되며, 서로 다른 언어로 인해 분리된 경우는 '+'가 부여된다.

그러나 심벌+한글의 경우 두 어절로 분리시 한글어절에 '+'가 부여되는데 경우에 따라서 부적절한 경우가 발생한다. 특히 인용부호나 괄호 다음에 '-'결합정보를 갖는 형태소가 올 수 있는데 예로 ["그 사람"]의 경우 [" +그 사람 "]으로 분리되어 어휘사전에 등록된 결합기호 '-'를 갖는 대명사 '그'와 매칭되지 못한다. 따라서 이와 같은 경우 결합기호 '*'가 부여되도록 하며 또한 줄바꿈문자에 의해 분리된 경우에도 '*'가 부여된다. 그 외의 경우는 '+'나 '-'만이 문장으로부터의 결합정보로 부여된다.

다음 표 3은 어절 및 형태소 분리시 입력문장으로부터 얻는 결합기호 발생위치 및 종류를 표시한 것이다.

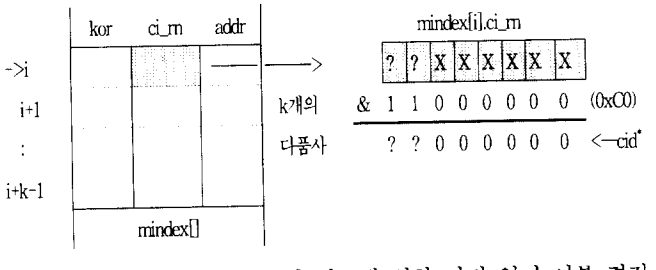
표 3. 결합기호 발생위치 및 종류
Table 3. Generating position and kind of conjunctive symbols.

입력문장	결합기호 발생위치	2차 결합기호
ъK	φ-K	※한글에만 결합기호 부여
ъ{C,E,N,S}	φ{C,E,N,S}	
X{C,E,N,S}	Xφ{C,E,N,S}	
SK	Sφ*K	
{K,C,E,N}K	{K,C,E,N}φ+K	
K1↵K2	K1φ*K2	K1φ-K2 / kn~K2 / K1~K2
{C,E,N,S}↵K	{C,E,N,S}φ*K	
X↵{C,E,N,S}	Xφ{C,E,N,S}	

※ {+,-,*}K -> {+,-,*}k1φ+k2φ+...+φ+kn
{ъ↵, ↵ъ, ъ↵ъ} -> ъ
ъ : {Space, Tab}ˆ φ : Space ↵ : LF or CR+LF
X : {K,C,E,N,S} K/C/E : Korean/Chinese/English word (k : Korean morphem)
N : Numeric S : Symbol

3. 줄바꿈문자 처리

각종 문서, 논문, 서적 등을 DB화할 경우 직접 키보드로 입력하거나 OCR에 의해 전자문서화 시키게 되는



cid*			cis		
기호	bit	값	*	-	+
*	0 0 0 0 0 0 0 0	0x00	○	○	○
-	0 1 0 0 0 0 0 0	0x40	○	○	×
+	1 0 0 0 0 0 0 0	0x80	○	×	○

○ : 읽기 ×:무시

그림 3. cis와 cid의 비교에 의한 사전 읽기 여부 결정
Fig. 3. Decision of access to dictionary by comparison cis with cid.

데 키보드로 입력한 경우 매 줄마다 마지막에 줄바꿈문자('\n')가 포함된 경우가 있으며 특히 OCR에 의한 경우는 각 줄마다 줄바꿈문자가 들어가 있어 각 줄의 마지막 어절과 다음 줄 처음 어절이 실제로 붙어야 하는지 띄어야 하는지를 전혀 알 수가 없게 된다. 영어의 경우는 단어 단위로 띄어쓰기를 하며 만약 하나의 단어를 다음 줄에 걸쳐 쓰고자 할 때는 반드시 hyphen(-)을 붙이게 되므로 이러한 문제가 없지만 한글의 경우 전혀 띄어쓰기 규칙의 영향을 받지 않고 음절단위로 끊어 다음 줄에 쓸 수 있다.

만약 붙여야 할 것을 떨어진 것으로 보고 분석하면 분석 실패할 확률이 높아지고 분석 성공했다 하더라도 그것은 오분석이 되며 반대로 떨어져 있어야 하는 것을 붙여야 할 것으로 보고 해석하면 오분석할 확률이 높아진다. 문제는 매 줄마다 이러한 경우가 발생하며 이를 무시하기에는 오분석이나 분석 실패할 확률 또한 높다는 점이다. 이를 해결하기 위해 결합정보를 줄바꿈문자 처리에도 적용한다.

한글 다음에 공백문자 없이 바로 줄바꿈문자가 오고 그 다음 또한 공백문자 없이 한글이 올 경우 즉, 결합기호로 '*'가 부여되는 경우의 문장형태는 3가지가 있으며 줄바꿈문자를 사이에 두고 나뉘어진 이 두 어절은 실제 결합기호 '- + ~'중 하나에 해당된다. 다음 표 4는 줄바꿈문자가 포함된 경우의 예이다.

표 4. 줄바꿈문자가 포함된 문장 예
Table 4. An example of sentences include a newline character.

형태	문장 예	결합기호	원래 의미	올바른 판독 예
결합	대마도↵ 도 우리땅	대마도 *도 -우리땅	+	대마도+도 -우리땅
결합	대↵ 마도도 우리땅	대 *마도도 -우리땅	~	대마도도 -우리땅
분리	대마도도↵ 우리땅	대마도도 *우리땅	-	대마도도 -우리땅
분리	대마도도␣↵ 우리땅	대마도도 -우리땅		※ ↵전후에 ␣있으면 무시

줄바꿈문자 다음의 어절 앞에 결합기호 '*'이 붙게 되는 경우 '*'는 어휘사전에서의 의미와는 약간의 차이가 있다. '*'을 어떻게 보며 어떻게 처리할 것인지는 다음과 같은 여러 가지 방법이 있을 수 있다.

① '*'

'*'를 그대로 적용한다. 즉, 앞 어절과 붙을 수도 떨어질 수도 있다고 보는 것으로 문장으로부터의 결합정보가 '*'이므로 어휘사전의 모든 동형이의어를 읽어온다.

해당 예 : 분석 *성공률(복합명사에서 단일명사 사이에 줄바꿈문자가 들어간 경우)

이 경우 '-'로 처리하면 단일명사 2개만 추출되지만 복합명사 결합규칙을 적용하면 해결된다. 또한 '+'나 '~'로 보고 처리해도 해결되므로 '*'만이 대안은 아니다. 또한 오분석확률이 가장 높아지기 때문에 처리대상에서 제외한다.

② '-'

'-'는 줄바꿈문자를 하나의 공백문자로 보는 것이다. 즉, 두 어절이 원래부터 떨어져 있는 것으로 간주한다. 따라서 우측 어절은 '-'라는 결합기호를 갖는다.

해당 예 : 돈을 조금만 *세어라(1)
한국에서의 *미적 기준(2)

그러나 실제 '+'나 '~'인 경우라면 좌, 우 어절중 분석실패할 확률이 높으며 성공하더라도 오분석이 되고 만다. 예로 '한국 *에서의 미적 기준'과 '돈을 조 *금만 세어라'의 경우 '*'은 실제 각각 '+', '~'가 되어야 하며 '-'으로 처리하면 분석실패한다.

③ '+'

'+'는 좌우어절이 줄바꿈문자에 의해 나뉘어졌다 하더라도 이를 하나의 어절로 보되 형태소를 구분하는 구분자(delimiter)로 보는 것이다.

해당 예 : 설계 *도 하나의 방법이다.

만약 '-'로 본다면 조사 '도'로 해석되지 못하므로 사전내용에 따라 분석실패나 오분석이 유발된다. '~'로 본다면 줄바꿈문자에 의해 분리되지 않은 정상 어절 상태와 같으므로 상관없으나 '+'로 볼 경우 분석이 용이한 경우와 분석이 실패할 경우가 있다. '~'로 볼 때 위 예는 '설계도'가 되며 이는 설계도(設計圖)와 설계+도(조사) 2가지 후보 패턴이 존재한다. 물론 이 경우는 정상어절의 경우에도 나타나는 애매성이지만 위 예에서 '+'로 본다면 '설계+도'의 분석결과를 얻을 수 있다. 하지만 '작업회 *의'의 경우는 잘못 판단하게 된다. 따라서 '*'를 '+'로 본다고 할 때 나타나는 문장의 의미에 따라서 그 결과가 옳을 수도 틀릴 수도 있으므로 굳이 이 방법을 선택할 필요가 없다.

④ '~'

'~'는 좌우어절이 실제 결합된 것으로 보고 하나의 어절로 간주한다. 이 경우는 분석 성공할 확률은 높아지지만 오분석할 확률 또한 높아진다.

해당 예 : 민족중 *흥의 역사적 사명

이 경우는 '~'이외의 다른 방법은 모두 분석실패가 되지만 '한국의 *사상가'의 경우 '~'로 보면 '한국/의사/상가'로 오분석할 수도 있다. 또한 ②의 예 (1)(2)의 경우 최장일치법 적용시 '만세', '의미적'이라는 형태소가 각각 추출될 수 있으므로 오분석이 유발된다.

이와 같이 4가지 방법의 장단점을 분석해본 결과, 분석성공률을 높이기 위해서는 '~'로 간주하고 분석하면 되지만 그만큼 오분석 확률도 높아지므로 본 논문에서는 분석실패할 경우가 많은 경우의 방법, 그러나 정확도는 그만큼 더 높은 방법으로 먼저 분석한 후 분석실패가 발생하면 2차 적용시 분석성공률이 높은 방법을 사용한다. 즉, '*'을 변환하는데 있어 '*,+'를 제외한 '~,-'로 2차 변환시 다음과 같은 기준으로 선택한다.

[기준 1] 앞, 뒤 어절 중 1음절인 경우 '~'로 보고 분석한다.

[기준 2] 그렇지 않은 경우 '-'로 보고 분석한다.

[기준 3] 기준 2 적용시 앞, 뒤 어느 한 어절이라도 분석실패하면 '~'로 보고 다시 분석한다.

```
MorphAnalysis(B)
{
  // A(어절 or 형태소) *B(어절)      A분석실패->어절, 성공->형태소
  //  (현재 분석완료)  (분석할 어절)
  X = 다음 분석할 어절
  IF X의 결합기호='*' AND A≠심벌
    // X는 B를 의미
    IF A분석성공 AND A길이>1자 AND B길이>1자
      IF Parse(B)=분석성공 RETURN
      DelNodes()  // B분석실패시 분석과정중 발생한 노드 삭제
    ENDIF
    // A분석실패한 경우, B분석실패한 경우, A 또는 B가 1글자인 경우
    X = A + B
  ENDIF
  IF Parse(X)=분석성공 RETURN
  DelNodes()
  미등록어절 기록
}
```

그림 4. 줄바꿈문자 처리 알고리듬
Fig. 4. The processing algorithm for newline character.

기준 1을 설정하게된 이유는 앞뒤 어절중 어느 하나가 1음절인 경우는 통계적으로 하나의 어절이 분리된 경우가 대부분이기 때문이다.

다음 그림 4는 줄바꿈문자 처리를 위한 알고리듬이며, 그림 5는 '*'을 '~'으로 보고 분석하는 경우의 형태 및 결합범위이다. 이때 '*'의 앞뒤어절을 결합해 재분석해야 하지만 앞어절이 분석성공해 여러 형태소로 분리되었을 때는 재분석 시간을 절약하기 위해 앞어절의 마지막 형태소와 뒤어절을 결합해 분석한다.(↓는 분석완료 위치를 나타낸다.)

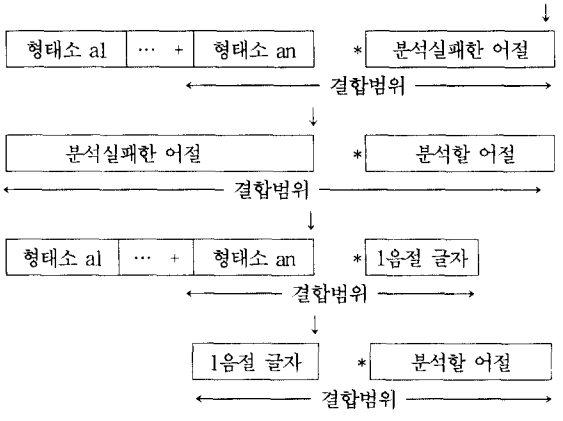


그림 5. *을 ~으로 처리하는 경우
Fig. 5. Cases of processing '*' as '~'.

4. 접사 추출

기존의 연구에서는 접두사, 접미사를 특별히 염두에 두지 않았는데 본 논문에서는 이러한 접사를 분리하여 처리함으로써 질의문에 포함될 수 있는 접사의 존재여부에 관련 없이 올바른 검색이 이루어지도록 한다.

1) 접사처리의 필요성

접사와 명사가 결합하여 하나의 단어로 굳어져 사용되거나 접사를 사용하지 않으면 뜻이 확연히 달라지는 경우라면 질의문과 문서에서 접사의 사용여부로 인한 불일치가 일어날 확률이 별로 없지만 접사와 명사가 결합하더라도 그 뜻이 크게 달라지지 않는 경우라면 질의문에는 접사가 사용되고 문서에서는 사용되지 않거나 혹은 그 반대인 경우가 있다. 이때는 검색실패가 되어 재현율의 저하를 초래하며 부분매칭을 하더라도 엉뚱한 단어와 매칭될 가능성이 있어 정확도를 떨어뜨리는 요인이 된다.

예를 들어 '가계약(假契約)'의 경우 질의어가 '가계(家系)'라면 '가계약'과 부분매칭되므로 엉뚱한 단어와 매

칭된다. 이를 방지하기 위해서는 '가계약'은 '가^계약'임을 알아야 한다. 이처럼 접사는 명사추출에 직접적인 연관이 있으며 대개 1음절인 관계로 다품사인 경우가 많다. 이 역시 결합정보를 사용하여 다품사 수를 줄인다.

2) 다품사에 의한 중의성

다품사에 대한 중의성으로 접두사간의 동형이의어와 접미사간의 동형이의어도 있지만 서로 다른 품사간의 동형이의어 즉 다품사 문제는 漢字를 한글로 표기함으로써 발생되며 이를 접사와 (불완전)명사, 접두사와 접미사, 접사와 기타간의 중의성으로 나누어 표 5에 보인다.

표 5. 접사와 관련된 다품사 예
Table 5. Examples of multiple parts of speech concerned with affix.

접사와 (불완전)명사	重^ 中(명사); ^果 科(명사); ^祖 條(불명) 조(명사); ^帽 모(명사,불명); ...
접두사와 접미사	假^ ^家; 輕^ ^鏡; 大^ ^代; 未^ ^美; 亞^ ^兒; 女^ ^餘; 兩^ ^洋; 一^ ^日; ...
접사와 기타	眞^ -진; 多^ -다; 固^ -고; 過^ -과; ^下 -하-; ^者 -자; ^人 -인; ...

접미사는 명사 뒤에 오며 또한 그 뒤에 기능어가 올 수 있는 위치특성으로 말미암아 접두사에 비해 많은 중의성을 유발한다. 이 중 접두사와 접미사간 동형이의어 관계에 있는 경우 표 6에 그 예를 보인다. 표 6의 접사간 중의성은 결합정보만으로 항상 간단히 결정할 수 있으며 접사와 명사의 중의성도 결합정보를 이용할 수 있다.

표 6. 동형이의어의 접사
Table 6. Affixes of same character.

가 각 강 개 견 경 고 공 과 군 금 급 노 내 단 담 당 대 되 막 미 민 별 복 새 생 선 소 쇠 수 시 실 아 암 애 양 여 옥 용 우 원 이 일 잔 장 저 전 정 제 종 주 짓 짝 차 처 초 치 토 통 항 해 호 ...

다품사에 관한 예 중 접미사가 조사 및 어미와의 중의성을 발생시키는 경우에 대해서는 [10]에 자세히 나와 있다. 접미사일 가능성은 앞에 오는 명사의 의미적 특성을 조사하면 알 수 있으며[10] 어근/어미/조사 등은 문법적 특성을 이용하면 관계없는 후보를 제외시킬 수 있다. 그러나 명사의 의미적 특성을 조사하기 위해서는 어휘사전의 모든 명사에 그 통사적, 의미적 특성에 대한 정보를 수록해야만 한다는 부담이 있으며 접미사와 기능어는 같은 결합정보를 가지므로 결합정보 처리대상에서 제외된다.

본 논문에서는 결합정보로 접사와 관련된 중의성 문제를 해결하도록 하며 표 7에 다른 품사와 동형이의어 관계에 있는 접사가 문장에서 '+'와 '-'형태로 각각 사용되었을 때 결합정보에 의해 선택되는 품사를 보이고 있다.

표 7. 결합정보에 의한 접사의 중의성 해결
Table 7. Solving multiple meaning of affix by the conjunctive information.

cis	p&s	p&(uc)n	s&(uc)n	p&f	s&f
+	s	(uc)n	s, (uc)n	f	s, f
-	p	p, (uc)n	(uc)n	p	

cis : 문장으로부터의 결합정보
p/s : 접두사/접미사, 사전 결합정보는 -/+
(uc)n : (불완전)명사, 이때 사전 결합정보는 모두 '*'라고 가정
f : 기능어, 사전 결합정보는 '+'

다음은 복합명사에 표 6의 접사가 포함된 경우이다. 먼저 단일명사일때 접사가 올 수 있는 위치는 다음과 같다.

p^N^s (p:prefix, N:Noun, s:suffix)

이로부터 도출되는, 문장으로부터의 결합정보는 접두사와 접미사가 각각 '-'와 '+'임을 알 수 있다. 따라서 이 결합정보로 p인지 s인지 알 수 있다. 2개의 명사가 공백 없이 결합된 복합어라면 접사가 올 수 있는 위치는 다음과 같다.

p_1^N_1^s_1/p_2^N_2^s_2

만약 'N_1aN_2'와 같은 경우에 접사 a가 표 6의 접사라면 문장으로부터 도출된 결합기호는 '+'가 되어 표 7에 의해 s로 간주된다. 하지만 실제로 a는 p일 수도 있으므로 이와 같은 패턴의 경우 결합정보만으로 품사중의성 문제를 해결할 수 없다. 즉, 어휘사전에 접두사의 결합정보로 '-'가 기록되어 있으므로 접두사의 가능성은 초기에 배제되어 분석오류의 가능성이 있기 때문이다. 본 시스템에서는 이와 같은 경우의 확률을 무시할 만한 것으로 보고 s로 보도록 한다.

Ⅲ. 실험결과 및 분석

1. 실험문서

실험에 사용된 분석대상의 문서는 모두 아무런 태그가 없는 비정형화된 ASCII파일로 '한겨레21' 기사와 석박사 졸업논문 초록 및 전문으로 총 479개의 파일로 구성하였다. 문서의 분야별 분포는 표 8과 같다.

표 8. 문서종류
Table 8. Kind of documents.

종류	한겨레21	석박사 졸업논문					계
		경제	국어	전산	수학	농업	
편수	310	33	36	33	32	35	479

2. 어휘사전

자체 제작한 어휘사전에는 약 10만 단어(10,1642 레코드)가 등록되어 있으며 한글학회의 '우리말 큰 사전'을 일부 참고로 하였다. 표 9는 어휘사전에 기록된 형태소의 n^+,n^-,n^*에 대한 통계로 평균 다품사 수는 예상대로 1음절의 경우가 1.8로 가장 많음을 알 수 있으며 2음절 이상은 다품사인 경우가 별로 없음을 알 수 있다. 본 논문의 목적은 최소한의 비용으로 다품사 수를 줄여 색인어를 추출하는데 있으므로 다품사가 많이 발생하는 1음절 중 다품사가 아닌 형태소를 제외한 1음절 다품사에 대한 통계도 함께 나타내었다. 1음절 다품사인 경우 하나의 형태소가 갖는 품사 수는 평균 약 2.8개임을 알 수 있다. 일반적으로 사전에서 명사가 차지하는 비율이 월등히 크고 명사는 대부분 결합정보로 '*'를 가지므로 n^*이 다른 경우에 비해 큰 수를 가진다. 그러나 다른 음절에 비해 기능어가 많은 1음절의 경우 역시 n^*의 비율이 높다. 그 이유는 1음절 다품사인 경우 342개중 명사가 264개이며 나머지 대부분은 불완전 명사로 띄어쓰기 오류를 허용하기 위해 가급적 '*'를 많이 부여하였기 때문이다.

한 형태소가 문장에서 '+,-'형태로 각각 사용되었을 때 다품사 감소율의 평균값은 표 2와 표 9로부터 구할 수 있다. 즉, 평균r_{1+}=평균n^+/a, 평균r_{1-}=평균n^-/a이므로 1음절인 경우 걸러지는 품사 수는 1.8000에서 각각 0.5757과 0.3700으로 약 32%(0.5757/1.8)와 21%(0.37/1.8)씩 감소함을, 1음절 다품사인 경우는 2.7722에서 각각 0.9525와 0.7373으로 약 34%(0.9525/2.7722)와 27%(0.7373/2.7722)씩 감소함을 추정할 수 있다.

표 9. 어휘사전의 n^+, n^-, n^*에 대한 통계
Table 9. A statistics of n^+, n^-, n^* in the dictionary.

음절 수	n	k	a	총n^+	총n^-	총n^*	평균n^+	평균n^-	평균n^*
1다품사	**876**	**316**	**2.7722**	**233**	**301**	**342**	**0.7373**	**0.9525**	**1.0823**
1	1,260	700	1.8000	259	403	598	0.3700	0.5757	0.8543
2	41,504	41,086	1.0102	126	1,753	39,625	0.0031	0.0427	0.9644
3	30,568	30,480	1.0029	46	3,003	27,519	0.0015	0.0985	0.9029
4	20,339	20,288	1.0025	10	4,306	16,023	0.0005	0.2122	0.7898
5	5,420	5,410	1.0018	1	2,092	3,327	0.0002	0.3867	0.6150
6	1,826	1,823	1.0016	0	821	1,005	0	0.4504	0.5513
7	399	399	1	0	60	339	0	0.1504	0.8496
8	103	103	1	0	6	97	0	0.0583	0.9417
9이상	44	44	1	0	0	44	0	0	1
총계	101,642	100,511	1.0113	442	12,515	88,685	0.0044	0.1245	0.8823

n : 레코드 수　　k : 유일한 형태소 수
a : 평균 다품사 수 = n/k

Input sentence : [정보검색 시스템에 관한 내용중 비정형화된 문서를 입력하면 HTML문서로 변환
환하는 시스템이 있다.]
Tokenized sentence(96) : 정보검색 시스템에 관한 내용중 비정형화된 문서를 입력하면 HTML +문서로 변 *환하는 시스템이 있다
morphem list --> (정보,0002)(+검색,0002)(시스템,0002)(+에,2008)(관한,0800)(내용,0002)(+중,4006)(비정형,0002)(+화,4003)(+된,0204)(문서,0002)(+를,0008)(입력,0002)(+하면,0200)(HTML,0002)(+문서,0002)(+로,0009)(변환,0002)(+하는,0200)(시스템,0002)(+이,045f)(있다,2000)
morphem list --> (정보,0002)(+검색,0002)(시스템,0002)(+에,0008)(관한,0800)(내용,0002)(+중,4002)(비정형,0002)(+화,4003)(+된,0200)(문서,0002)(+를,0008)(입력,0002)(+하면,0200)(HTML,0002)(+문서,0002)(+로,0009)(변환,0002)(+하는,0200)(시스템,0002)(+이,044b)(있다,2000)
morphem list --> (정보,0002)(+검색,0002)(시스템,0002)(+에,0008)(관한,0800)(내용,0002)(+중,4000)(비정형,0002)(+화,4001)(+된,0200)(문서,0002)(+를,0008)(입력,0002)(+하면,0200)(HTML,0002)(+문서,0002)(+로,0008)(변환,0002)(+하는,0200)(시스템,0002)(+이,0448)(있다,2000)
[1]Selected keywords from the result of morphological analysis.
정보/검색P1시스템 내용 비정형(^화) 문서P4입력 HTML/문서변환P6시스템

그림 6. 결과화면
Fig. 6. Result screen

3. 문장 분석결과

실험은 486 PC에서 Linux 운영체제를 기반으로 하

여 C언어로 구현한 색인기로 하였다. 형태소분석시 분석과 함께 통계자료 파일도 자동 생성되도록 하였으며 수행시간은 약 10분이었다.

그림 6은 하나의 문장을 형태소분석하는 중간 과정을 화면 출력한 내용이다.

그림 6의 Input sentence는 파일로부터 읽은 하나의 문장이며 Tokenized sentence는 공백문자로 분리되어 있는 원래의 입력문장을 본 논문 II-2의 어절분리 과정을 거치면서 결합기호가 부여된 후의 문장으로 파일로부터의 입력과 어절분리는 1 pass로 이루어진다. 결합기호가 '-'인 경우 화면상에서는 표시되지 않는다. morphem list는 형태소 분석된 결과를 나타내는데 입력문장으로부터 부여된 결합기호, 형태소, 중첩(bitwise OR)된 16진수 품사 값으로 표시되어 있다. 첫 번째 형태소 리스트는 결합정보를 사용하지 않았을 때이며 두 번째 리스트는 사용하였을 때, 세 번째는 형태소분석후 2차로 다품사 수를 줄였을 때의 예이다. 이 경우 '에, 중,화,된,로,이'의 품사 수 변화를 중첩된 16진수 품사값의 변화를 통해 알 수 있는데 특히 '+이'의 경우 품사값이 045f > 044b > 0448로 변화되는 것을 볼 수 있다. 정의된 품사값은 그림 7과 같다.

```
#define SUF       0x0001   //접미사
#define NOUN      0x0002   //명사
#define PRE       0x0004   //접두사
        :
#define IDIOM     0x0800   //관용구
#define SYMBOL    0x1000   //심벌
#define ETC       0x2000   //기타
#define STOPWD    0x4000   //불용어
#define UNREG     0x8000   //미등록어
```

그림 7. 품사정의
Fig. 7. A definition about part of speech.

4. 실험결과 분석 및 평가

제안한 알고리듬을 적용한 결과를 평가하기 위해 기존의 시스템과 비교하여야 하나 현 상황에서는 자동색인과 관련하여 공개된 통합시스템이 별로 없어 일반 텍스트문서에 태그를 붙인 KTSet과 일부 형태소분석기가 있을 뿐이다. 특히 형태소분석기는 전자사전의 구조 및 검색방법과 밀접한 관계에 있으나 지적재산권 문제로 인해 전자사전의 완전한 내용공개는 쉽지 않은 문제이다.

각종 문법규칙과 의미정보, 태깅된 코퍼스 등이 필요한 시스템과 그렇지 않은 본 시스템과의 비교가 적절하지 않기 때문에, 제안한 알고리듬의 효율에 영향을 미치는 전자사전을 구축한 후 표 9의 자체 통계 조사로부터 품사후보를 몇 % 제거할 수 있는지를 추정할 수 있었다. 또한 표 10은 각 문서의 형태소분석후 생성된 통계자료를 모두 더한 전체 문서에 대한 통계 및 분석결과이다.

형태소분석 시간은 총 10분으로 486 PC에서 약 1,000 [형태소/sec] 의 분석속도를 보였으며 실패어절은 6,976어절로 약 98.16%의 분석성공률을 보였다. 참조횟수는 2차 품사후보 제거과정을 거치기 전의 결과로 인덱스 참조횟수가 아닌 디스크사전 참조횟수를 말하며, 이 중 다품사에 관해서는 결합정보를 사용하지 않았을 때 744,631회, 결합정보를 사용하였을 때는 572,714회로 171,917회 줄어 결합정보 사용시 23.1% 감소함을 보였다.

다품사 수를 감소시키는데 있어 그 정확도는 입력문장의 띄어쓰기 형태에 따라 좌우되며 실수하기 쉬운 띄어쓰기 형태의 형태소는 '*'를 부여하였으므로 일부러 붙여써야 할 것을 띄어쓰지 않는 한 품사를 잘못 제거하지는 않는다. 따라서 잘못 제거한 품사에 대한 조사는 하지 않았다.

Ⅳ. 결 론

본 논문에서는 색인기를 위한 형태소분석시 입력문장으로부터 자동획득한 결합정보를 이용하여 다품사수를 줄이는 방법을 제안하였으며 줄바꿈문자로 인해 어절이 강제로 분리된 경우에도 결합정보의 관점에서 보고 이를 해결할 수 있었다. 또한 색인어와 직접적인 연관성이 있는 접두사, 접미사의 중의성 문제를 살펴보고

표 10. 분석결과
Table 10. Analyzed result.

문서 크기		분석후 통계					다품사에 대한 참조수		
문장 수	어절 수	형태소 수	명사 수	접사 수	참조횟수	실패어절	사용안함	사용	감소율
80,998	379,187	596,059	246,809	73,718	842,607	6,976	744,631	572,714	23.1%

보고 결합정보로 품사후보수를 줄일 수 있음을 보였다.

제안한 방법은 사전구축시 간단한 결합정보만 부여하면 분석 가능성이 없는 품사 후보를 미리 제거하기 때문에 매우 효율적이다. 즉, 복잡한 품사간 또는 형태소와의 규칙이 필요 없고 방대한 의미정보가 필요한 사전 또는 태깅된 코퍼스의 구축에 따르는 어려움을 피할 수 있었으며 코퍼스를 가지고 하는 학습과정 또한 필요 없이 형태소분석시 즉시 문장으로부터 결합정보를 자동획득할 수 있었다.

제안한 알고리듬의 다품사 수 감소율을 보이기 위해 여러 분야에 걸친 문서를 대상으로 결합정보를 사용한 경우와 그렇지 않은 경우를 각각 실험해 본 결과, 결합정보를 사용한 경우가 사용하지 않은 경우에 비해 다품사 수가 23.1% 감소되었다. 이를 띄어쓰기 오류가 없는 문서에 적용한다면 더 많이 감소되며 특히 1음절 다품사의 경우는 III-2에서 언급했듯이 더 높은 비율로 감소된다. 이 방법은 문법규칙이나 사전구축, 통사분석에 들어가는 추가비용없이 빠른 색인시스템을 구현하고자 할 때, 그리고 최장일치법 적용시 발생하는 다품사 수 감소방안으로 사용될 수 있다.

형태소분석후 우결합정보와 품사결합정보를 이용해 2차로 다품사 수를 줄이도록 하였으나 추후 이에 대해 최소한의 비용으로 품사 수를 더 줄일 수 있는 연구가 이루어져야 할 것이다. 아울러 접사와 결합정보가 같은 품사와의 중의성 문제 해결을 위한 접사-명사 관계사전에 대한 연구가 필요하다.

참 고 문 헌

[1] Frakes, "Information Retrieval," *Prentice Hall*, 1992.

[2] 정영미, "정보검색론", 구미무역, 1993

[3] 강승식, "한국어 형태소 분석을 위한 복합 명사의 인식 방법", 한국인지과학회 춘계 학술발표논문집, pp. 175-189, 1993

[4] 김판구, 조유근, "상호정보에 기반한 한국어 텍스트의 복합어 자동색인", 한국정보과학회 논문지, 1994. 7

[5] 윤보현, 임희석, 임해창, "통계정보를 이용한 한국어 복합명사의 분석방법", 한국정보과학회 봄 학술발표논문집, pp. 925-928, 1995

[6] 이현아, 홍남희, 이종혁, 이근배, "한국어 형태소 구조규칙에 기반한 색인 시스템의 구현", 한국정보과학회 봄 학술발표논문집, pp. 933-936, 1995

[7] 양재형, "공기 유사성을 이용한 한국어 명사구 접속의 구조적 모호성 해결", 정보과학회논문지, 제23권, 제3호, B권, pp. 311-321, 1996. 3

[8] 임해창, 임희석, 이상주, 김진동, "자연어 처리를 위한 품사 태깅 시스템의 고찰", 정보과학회지, 제14권, 제7호, pp. 36-56, 1996

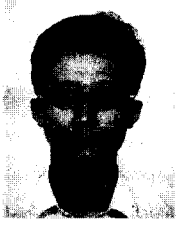
[9] 김재훈, 김길창, "언어지식을 이용한 형태소 해석의 모호성 축소", '96 제8회 한글 및 한국어 정보처리 학술대회 논문집, pp. 231-234, 1996

[10] 남윤진, 옥철영, "말뭉치 분석에 기반한 명사파생접미사의 사전정보 구축", 정보과학회논문지, 제23권, 제4호, B권, pp. 389-401, 1996. 4

[11] 서창덕,신청식,한기태,김진웅,임인철, "한-일 기계번역을 위한 조사 및 관용구 처리 알고리즘", 전자공학회 추계학술발표논문집, pp. 579-582, 1989

[12] 최기선, "한국어 정보검색", 정보과학회지, 제12권, 제8호, pp. 24-32, 1994. 9

저 자 소 개



徐 昌 德(正會員)

1964년 5월 7일생. 1988년 2월 한양대학교 전자공학과 졸업(공학사). 1990년 8월 한양대학교 대학원 전자공학과 졸업(공학석사). 1996년 2월 한양대학교 대학원 전자공학과 박사과정 수료. 관심분야는 한글 자연어처리, 정보검색, 자동색인, 하이퍼미디어

林 寅 七(正會員) 第 30卷 B編 第 2號 參照

현재 한양대학교 전자공학과 교수